**Vivekkumar Singh**

14305 Wedgewood Ct Tampa, FL- 33613 |Phone: 8136062975 | [viveks2692@gmail.com](mailto:viveks2692@gmail.com) | [Linkedin](https://www.linkedin.com/in/vivekkumarusf2020/) | [Github](https://github.com/Vivekkumar260892): Vivekkumar260892

**SUMMARY**

Over 4 years of comprehensive work experience in Data Engineering, Marketing Analytics and Business Intelligence in banking and retail domains. Ability to solve complex business problems using ETL, Data Mining, Machine Learning & Data Warehousing concepts.

**EDUCATION**

* MS in Business Analytics and Information Systems, University of South Florida ***(Jan 2019-May2020)***
* BE in Computer Engineering, University of Mumbai ***(Jul 2011- Jul 2015)***
* Certifications**:** [Google Digital Unlocked](https://drive.google.com/open?id=1hk-yppx7N8wtPaYNUsswoy-JoMxn-ktL), [AWS certified Cloud Practitioner](https://www.youracclaim.com/earner/earned/badge/916e6481-cbff-4165-982c-d460209cd2e1), [Tableau Data Scientist](https://www.youracclaim.com/badges/8e401080-061f-4a25-a041-c27f5f333b1f)

**WORK EXPERIENCE**

***Assistant Decision Sciences at University of South Florida: (Jun 2019- May 2020)***

* Performed Data **Cleaning**, Data **Manipulation** and Data **Integration** to maintain Data Quality for creating various assignments.
* Created **Tableau** dashboards to explain variation in success **Metrics** and **Time Series Analysis** toundergraduate students.
* Automated assignment grading process using **Excel VBA (Macros)** and **MySQL** maintaining accuracy and saving **~ 75%** of time.

***Data Engineer at Cartesian Consulting: (Mar 2018- Dec 2018*)**

* Developed pipelines for **ETL** (Extract, Transform, Load) using **MySQL**, **Python**, **Airflow** and **AWS S3** foracquiringaPOCproject.
* Extracted data from streaming pipelines using **Flume** and **Kafka** and processed using **Spark** Structured Streaming.
* Predicted sales by **time series forecasting** in **Python** using **neural networks, ARIMAX** and **Prophet** for inventory management by eliminating understocking and reducing overstocking by 56%.
* Applied **K-means clustering** in **Python** for **segmentation** of customers comparing it with **RFM** (Recency, Frequency and Monetary Value) analysis for improved campaign targeting.
* Developed **dimensional** **data** **models** and **data** **warehouse** adhering to integrity and **normalization** rules to support campaign **data** **mart** and customer one view for marketing campaigns. Wrote **complex** **SQL** queries (multiple joins, CTE’s, subqueries).
* Generated visualizations using **Tableau** toanalyze marketing **metrics** for making recommendations and supply chain analysis.

***Data Engineer at Hansa Cequity:* (*Mar 2016- Sep 2017*)**

* Identified probable customer churn using **Classification Models** in **Python** like **Decision Trees** and achieved a recall of 84%.
* Worked with the **Apache** **Spark** Framework for customer analytics using **Spark** **SQL** queries on large scale datasets for developing flawless **CRM** (customer relationship management) campaigns and deployed them through multiple channels.
* Pre-processed structured and unstructured data from different sources (**RDBMS**, **Hadoop**) using **Hive**, **SQL**, **Sqoop**.
* Developed an **ETL** pipeline for loading and wrangling data from **SSIS** to **IBM UNICA** for daily campaign workflows using **data** **warehousing** concepts.
* Leveraged Amazon Web Services like AWS **EC2**, **EMR**, **S3** and **Redshift** to pull data for various campaigns.
* Implemented **A/B testing**, **t-tests** and **ANOVA** strategies and analyzed impact on KPIs with understanding of **Project Management** and **Agile** methodologies. Validated **ROC** curves and **confusion** **matrices** for hyperparameter tuning.
* Created comprehensive **dashboards** in **Tableau** for measuring campaign performance and funnel reports.

***Business Intelligence Engineer at Intelenet Global Services:* (*Sep 2015- Feb 2016*)**

* Optimized complex **SQL** scripts for quality checking of projects and populating output tables for deployment using **SSIS**.
* Built **KPIs** and **Regression** models to predict **customer life-time value**, enhance propensity and scoring attributes.
* Automated hourly status report saving **10 man-hours/week**, thus decreasing response time for fixes and campaign failures.
* Accurately extracted insights and created dashboards using **Tableau,** **Excel VBA (Macros)**, **pivot** **tables** and **slicers**.
* Formulated ad-hoc reports based on requirements gathered from various stake holders using **JIRA** to provide solutions.

**PROJECTS**

***Anti-Money Laundering AML Model (Language/Tools- Python, Jupyter Notebook):***

* Applied models like Logistic Regression and Gradient Boosting to classify the customers as suspicious or not suspicious by data manipulation, feature engineering and variable selection techniques.
* Used various correlation tests and cross validation to reduce overfitting and achieved a recall (sensitivity) of 82% for classifying suspicious customers.

***Customer churn classification using different models (Tools- SAS Enterprise Miner):***

* Classified customers probable to churn after performing data cleaning, one-hot encoding and dimensionality reduction using PCA.
* Applied stratified sampling to ensure equal proportions of classes in the training set and learned how to save and deploy ML models in SAS Enterprise Miner.

***Topic Modeling on 20-Newsgroups dataset (Language/Tools- Python, Jupyter Notebook):***

* Imported the corpus from the json source using pandas read json command and created bigrams, trigrams using term frequency inverse term frequency (TF-IDF) vectorizer.
* Built Latent Dirichlet Allocation (LDA) model to discover different topics hidden in the corpus and visualized them using the interactive chart in pyLDAvis package.

***Predicting work authorization status (Language/Tools- Pyspark, Spark SQL, Databricks Notebook):***

* Used the Random forest ensemble from Pyspark ml library in order to predict lottery selection on the Kaggle dataset (3 million records) with an accuracy of 91% on the Databricks platform.
* Used chi-square correlation test and Spark SQL to analyze, manipulate, clean and merge data before building a pipeline to pre-process both train, test sets by applying the indexer, assembler and one hot encoder.

**SKILLS**

* ***Programming:*** SQL, Python, R, SAS, Pyspark, HTML, C#, Excel VBA (Macros).
* ***Big Data Ecosystem***: Spark, Hadoop, MapReduce, Hive, Pig, Kafka, Flume.
* ***Cloud Technologies***: AWS (S3, EC2, Lambda, Athena, RDS, Redshift, EMR).
* ***Tools:*** Tableau, Power BI, Azure ML, RStudio, Jupyter Notebook, SAS E-Miner, SAS CI, IBM-Unica, SSIS, MS Office, JIRA.
* ***Libraries*:** Numpy, Pandas, Matplotlib, Seaborn, Scikit-Learn, Keras, Nltk, Gensim, Scipy, Beautiful Soup.